Human-Centered Responsible AI: Current & Future Trends

Special Interest Group
CHI '23

Instructions

Please duplicate the template slide (next slide) and modify it as needed.

TEMPLATE Group name

Situation:

The framing of the important, recent context the audience already knows and accepts as fact.

Complication:

The reason the situation requires action.

Resolution:

The action required to solve a problem (or capture an opportunity)

Tools for ensuring responsibility in AI-based applications

Situation:
Application developers create applications that use AI
Complication:
It is not straightforward to ensure that the applications meet responsibility criteria
Resolution:
Can we create “responsibility linting tools” for design tools?
-- We need to crystallize useful models for this purpose
-- Distinguish between being responsible and showing others that you’re responsible
-- People tend not to use large guidelines sets, so they should be made aware of issues as they go along
-- It takes a group to make responsible AI, since members can represent different perspectives and critique each other.

Tools & Frameworks - ONLINE

Situation:
AI Frameworks with many different stakeholders in different context and domain.

Complication:
Tools are not realistic in real-world settings. Too complex and it takes too much time to get used in the new tools.

Resolution:
More cases studies of use needed to refine frameworks
Literature reviews of recent tools and frameworks might needed.
Types: fairness, accountability, transparency, control, human-in-the-loop

Tools & Frameworks: A/B Testing to bridge HCI & AI

Situation:
HCI is concerning with identifying more vs less effective interventions and designs, and making *changes*- e.g. to instructions, interfaces.
Complication:
Most ML/AI techniques focus on learning from huge data sets to make predictions, but don't provide statistical or other techniques for identifying which of a set of designs is effective, in which context.
Resolution:
Provide tools/frameworks for enabling designers to use Reinforcement Learning/Bandit Algorithms for adaptive A/B testing: Humans formulate micro-design alternatives as conditions/arms, and algorithms.
http://tiny.cc/moocletpaper is one framework.

Tools & Frameworks for End Users

Situation:
NHS designed an AI system to predict risk of COPD (a potentially fatal lung disease). NHS technologists report that they've used fairness tools (e.g. Fairness360) and tried to engage impacted patients in conversations about model function and fairness, but patients expressed "disinterest" in participating. We're a 3rd party evaluator being asked to weigh in if it's "Ethical AI".

Complications:

- We know people want a say in their own healthcare and suspect that the problem is that impacted people have not be engaged appropriately.
- System developers feel like they've done enough to consider AI fairness.
- Advising clinicians feel like this is a simple case of evidence-based medicine and outcomes.

Resolution:

- Look beyond just one "fairness" framework
- Deny the roll out of the tool until patients (and other stakeholders) are adequately involved
- Mandate use of HCI tools (remind them tools exist, dont have to recreate the wheel)

Use cases I

Situation: using AI in education and/or academia for authoring content

Complication:

- Risk of losing ideas, intellectual property
- Impact on identity
- Potential for exacerbating digital divide in awareness of AI;
- Homogenization of generated content - often Western, English
- Is it possible to have some standards of quality? May need to be industry-specific
- Risk of de-skilling, over-reliance on automated tools
- Lack of educational development and skill development - fairness implications
- Chatbots / LLMs replacing social interactions
- Impacts on our self-efficacy

Resolution:

- Could be an opportunity to educate students differently, rethink education
- More opportunities for humans to be reflective on what is unique about us
- Public education campaign around what these tools are, (e.g., hand-washing)
- We have models for human-human interaction already that we can draw on

Tools & Frameworks: A/B Testing to bridge HCI & AI

Situation:

HCI is concerning with identifying more vs less effective interventions and designs, and making *changes*- e.g. to instructions, interfaces.

Complication:

Most ML/AI techniques focus on prediction and learning from

Resolution:

The action required to solve a problem (or capture an opportunity)

Fairness, power dynamics, biases, diversity & beyond

Situation:

Use case: Higher Education

- Algorithms replacing students self-identifying & teachers or counselors identifying students who need support
- Retention replaced wellness as motivation

Complication:

- Metrics used not proven to actually indicate students’ engagement
- Actions prompted focused solely on student intervention, never a review of systems (LMS)
- Multiple stakeholders, different personal characteristics, changing environments - adjustments to interventions offered don’t consider these

Resolution:

Research questions:

- How do algorithms’ flags reflect students, teachers & counselors’ perception of student engagement?
- How can students, teachers & counselors be included in the design of the systems (participatory design)?
- Can automated systems be used to look out for student wellness, rather than simply retention?
- How can the chain of responsibility/accountability be made visible?

Fairness, power dynamics, biases, diversity & beyond - ONLINE

Situation:
The problem of who is responsible for the content and decisions of AI when the law has not caught up to the technical situation, is currently very murky.

Complication:
It takes a long time to craft laws, and there are many countries with different cultural context. These laws have not caught up to the realities of the tools. And the question of who are the laws protecting: business, users, creators, etc. can be in conflict.

Resolution:
Recognizing the difference between **legal, ethical and socio-cultural aspects** of fairness. The **multiplicity and change of roles** (user can be a source of data, and/or a target of a decision, and/or a receiver of an artifact for example) and **contexts** of data (the original purpose of museum collections for example is to educate, not to train AI) all need to be considered.

A Human-Centered Approach to Tackling Hallucinating LLMs

Situation:
LLMs will impact a plethora of applications.

Complication:
However, LLMs fail in being grounded and, at times, they can give perfectly sound yet false answers (hallucinations like “Why eating glasses is healthy”).

Resolution:
To tackle hallucination, we propose new ways of making LLMs more explainable by tapping into user feedbacks (through, e.g., new gamification approaches)

Harms and risks - face-to-face

Situation:

- We cannot control what happens to a photo that may include you
- Technological measures to avoid misuse of services

Complication:

- Bad actors can access said image that includes you and enact malicious intent
- Misuse of a systems for malicious/unintended purposes

Resolution:

- Develop perceptual designing to prepare for potential misusage
- Technical measures to lock in the system in case of misusage

Theory

Situation:
No established or widely used theories used in RAI
Complication:
What is a theory (e.g., framework, process, model)? What are different ways HCI use theories (e.g. recipe for design, evaluation, prediction)
Diverse places to use theories in RAI

Psychology theories to inform design, understand the role of technology (e.g. mental model, self-determination theory, sense-making, affordance)

Normative lens (social, ethical, political)

Process frameworks (e.g. double diamond)

How to take theories into RAI practices?
Resolution:
Unpack the stages of RAI, i.e. ideation and evaluation stages
Develop RAI native theories

Harms and Risks - Online (Vitak reporting out)

Situation:

- Lack of clear frameworks/definitions guiding design, implementation, and regulations for new AI.
- We design and build systems as if they are simple *technical* problems
- We design and build systems as if our data are perfect
- We assume that everyone has the same understanding of a data concept
- Lack of meaningful consent.

Complication:

- Lack of regulatory framework leads to “wild west”
- Cross-cultural considerations (e.g., different norms, values)
- Overreliance, overconfidence, and altering environments
- “Who” questions: Who designs for whom? Who is not heard? Whose needs are not considered, and who is not considering those needs? → Make the (multiple, diverse) people in the situation legible to one another.
- Data are messy, but our methods hide the messiness. We don’t know what might be wrong. We don’t flag uncertainties

Resolution:

- Make the (multiple, diverse) people in the situation legible to one another.
- Greater focus on policy making (e.g., see EU’s AI Act)
- Preserve data provenance to allow to re-inspection of earlier data-work manipulations and implied decisions
- Unpack assumptions underlying the design, implementation, and regulation
- Construct validity

Specific Use cases I : Human - AI teams (Online)

Scenario: Everyday task: "E-Mail"

Complication

scan beforehand about dangerous messages (content & attachments) – phishing attacks screening

user control on human-AI decision making

Recommendation made during the text for email trained on other human emails

Human - AI team – control (levels) is dynamic – OTP emails – levels of autonomy

why user decide some email is dangerous? -- AI statistics , human decision

Resolution: AI provides post- hoc explanations – statistical measures, human comes up with solutions